

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平10-247093

(43)公開日 平成10年(1998)9月14日

(51) Int.Cl.<sup>6</sup>

識別記号

FI

|            |      |
|------------|------|
| G 1 0 L    | 3/00 |
|            | 7/02 |
|            | 9/00 |
| // H 0 4 N | 7/24 |

**5 3 1**

|         |      |
|---------|------|
| G 1 0 L | 3/00 |
|         | 7/02 |
|         | 9/00 |
| H 0 4 N | 7/13 |

5 3 1 N  
A  
D  
Z

審査請求 未請求 請求項の数 8 FD (全 9 頁)

(21)出願番号

**特願平9-65595**

(22) 出願日

平成9年(1997)3月5日

(71)出願人 000001214

国際電信電話株式会社

東京都新宿区西新宿2丁目3番2号

(72)発明者 氏原 清乃

東京都新宿区西新宿2丁目3番2号 国際  
電信電話株式会社内

(72)発明者 中島 康之

東京都新宿区西新宿2丁目3番2号 国際  
電信電話株式会社内

(72)発明者 米山 曉夫

東京都新宿区西新宿2丁目3番2号 国際  
電信電話株式会社内

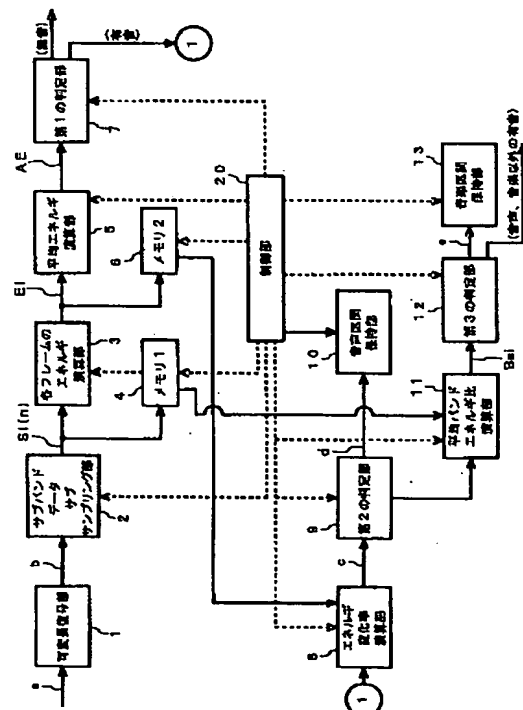
(74)代理人 弁理士 田中 香樹 (外1名)

(54) 【発明の名称】 オーディオ情報分類装置

(57) 【要約】

【課題】 音楽区間のみならず音声区間の検出もできるオーディオ情報分類装置を提供することにある。

【解決手段】 第1に、入力されたオーディオ情報から1秒毎の周波数データを抽出する。第2に、平均エネルギー演算部5は1秒間の平均エネルギーを求め、1秒単位で無音区間と有音区間の判定を行う。第3に有音区間であると判定された区間に対して、エネルギー変化率演算部8にて1秒毎のエネルギー変化率を求め、音声区間を抽出し、音声区間保持部10にそのタイムコードを入力する。第4に、音声区間と判定されなかった区間に対して、平均バンドエネルギー比演算部11にて1秒毎の平均バンドエネルギー比を求め、音楽区間を抽出し、音楽区間保持部13にそのタイムコードを入力する。



## 【特許請求の範囲】

【請求項1】 オーディオ情報から音声区間と音楽区間を分類するオーディオ情報分類装置において、  
入力されたオーディオ情報から単位時間ごとの周波数データを抽出するオーディオ周波数データ抽出手段と、  
抽出した単位時間ごとの周波数データを用いて、その区間が無音区間か有音区間かを判定し有音区間のみを抽出する無音／有音判定手段と、  
有音区間と判定された区間が音声であるか否かを判定する音声区間抽出手段と、 有音と判定された区間が音楽であるか否かを判定する音楽区間抽出手段とを具備したことを特徴とするオーディオ情報分類装置。

【請求項2】 請求項1のオーディオ情報分類装置において、

前記オーディオ周波数データ抽出手段によって抽出される単位時間ごとの周波数データは、入力されたオーディオ情報がMPEG符号化データである場合、単位時間分のMPEG符号化データの各フレームの先頭にあるサブバンドデータであることを特徴とするオーディオ情報分類装置。

【請求項3】 請求項1のオーディオ情報分類装置において、

前記無音／有音判定手段は、前記オーディオ周波数データ抽出手段により抽出された単位時間分の周波数データを用いて単位時間の平均エネルギーを求め、該平均エネルギーの大きさにより無音／有音区間を判定することを特徴とするオーディオ情報分類装置。

【請求項4】 請求項3のオーディオ情報分類装置において、

前記無音／有音判定手段は、入力されたオーディオ情報がMPEG符号化データである場合、単位時間の平均エネルギーは、MPEG符号化データの各フレームのサブバンドデータから求めたエネルギーの単位時間における総和であることを特徴とするオーディオ情報分類装置。

【請求項5】 請求項1のオーディオ情報分類装置において、

前記音声区間抽出手段は、前記オーディオ周波数データ抽出手段により抽出した単位時間ごとの周波数データからエネルギー変化率を求め、該エネルギー変化率の大きさにより、音声区間を抽出することを特徴とするオーディオ情報分類装置。

【請求項6】 請求項5のオーディオ情報分類装置において、

前記音声区間抽出手段は、入力されたオーディオ情報がMPEG符号化データである場合、エネルギー変化率は、MPEG符号化データのサブバンドデータから求めた隣り合うフレームの2つのエネルギーの比の単位時間における総和であることを特徴とするオーディオ情報分類装置。

【請求項7】 請求項1のオーディオ情報分類装置にお

いて、

前記音楽区間抽出手段は、前記オーディオ周波数データ抽出手段により抽出した単位時間ごとの周波数データから平均バンドエネルギー比を求め、該平均バンドエネルギー比から音楽区間を抽出することを特徴とするオーディオ情報分類装置。

【請求項8】 請求項7のオーディオ情報分類装置において、

前記音楽区間抽出手段は、入力されたオーディオ情報がMPEG符号化データである場合、平均バンドエネルギー比は、MPEG符号化データのサブバンドデータの全データに対する低周波帯域のサブバンドデータの割合であることを特徴とするオーディオ情報分類装置。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】本発明はオーディオ情報の分類装置に関し、特に、符号化されていない元のままのオーディオ情報あるいは符号化されたオーディオ情報から音声区間と音楽区間を分類できるオーディオ情報の分類装置に関する。

## 【0002】

【従来の技術】オーディオに関する研究は、今までは、周波数解析等を用いて計算機上に取り込まれた音声は何であるかを認識する音声認識や、調音パラメータ等によって機械的に音声を合成する音声合成の分野に関する研究が主流であり、オーディオをタイトルや内容によって分類するインデクシングに関する研究はまだ少ない。インデクシングに関する研究報告としては、例えば、南、阿久津らの“大量映像に対する効率的アクセスインターフェース”、ITE Technical Report Vol.19, No.7, pp.1-6のように音楽区間を検出し、その情報を用いて音楽が流れている動画をインデクシングするものがある。

## 【0003】

【発明が解決しようとする課題】しかしながら、この研究報告は、音声区間の検出に関しては何ら触れておらず、音声区間を検出することができないという問題がある。また、該研究報告は、音楽区間の検出に関しても、周波数スペクトルのピークをLPCケプストラムによって検出し、その平均持続時間を用いて音楽区間を検出しているため、圧縮符号化されたオーディオ情報からの検出は不可能であるという問題がある。

【0004】本発明の目的は、前記した従来技術の問題点に鑑み、音楽区間のみならず音声区間の検出もできるオーディオ情報分類装置を提供することにある。他の目的は、符号化されていないオーディオ情報および圧縮符号化されたオーディオ符号化データの両方でオーディオ情報を音楽区間と音声区間に分類することを可能にするオーディオ情報分類装置を提供することを目的とする。

## 【0005】

【課題を解決するための手段】前記目的を達成するため

10

20

30

40

50

に、本発明は、オーディオ情報から音声区間と音楽区間を分類するオーディオ情報分類装置において、入力されたオーディオ情報から単位時間ごとの周波数データを抽出するオーディオ周波数データ抽出手段と、抽出した単位時間ごとの周波数データを用いて、その区間が無音区間か有音区間かを判定し有音区間のみを抽出する無音／有音判定手段と、有音区間と判定された区間が音声であるか否かを判定する音声区間抽出手段と、有音と判定された区間が音楽であるか否かを判定する音楽区間抽出手段とを具備した点に第1の特徴がある。また、本発明は、入力されたオーディオ情報がMPEG符号化データであつても、符号化データ上でオーディオ情報を音声区間と音楽区間に分類できるようにした点に第2の特徴がある。

【0006】本発明によれば、符号化されていない元のままのオーディオ情報、あるいは符号化されたオーディオ情報のいずれからも、簡単かつ高速で、音声区間と音楽区間を分類することができるようになる。

【0007】

【発明の実施の形態】以下に、図面を参照して、本発明を詳細に説明する。図1は本発明の一実施形態の構成を示すブロック図、図2、図3は、図1のシステムの動作、特に制御部20の動作の概要を表わすフローチャートである。この実施形態は、動画像および音声符号化の国際標準方式であるMPEG1(ISO/IEC 11172)により圧縮された音声符号化データを用いて音声、音楽を分類するものであるが、本発明はこれに限定されるものではない。

【0008】以下に、図1および図2、図3を参照して、本実施形態の構成と動作を説明する。図1に示されているように、圧縮符号化されたオーディオの符号化データaは、可変長復号部1に入力される。ここで、該圧縮符号化されたオーディオの符号化データの構造について、MPEG1を例にして図4を参照して説明する。MPEG1では、図示されているように、元のオーディオ信号pからサンプリングした512個のPCMサンプルPをサブバンド符号化して32個のサブバンドデータP<sub>i</sub>(n)(n=0, 1, ..., 31)を作り、それを時間的にサンプルをずらしながら36回繰り返して、合計1152個のサブバンドデータを1フレームの符号化データQとしている。

【0009】前記した構造の符号化データQが前記可変長復号部1に連続して入力してくると、該可変長復号部1はこれを各フレームのサブバンドデータに復号し、サブバンドデータサブサンプリング部2に出力する。いま、ある単位時間を1秒とすると、該1秒は図5のaのように38フレームから構成されているので、可変長復号部1は1秒分の符号化データに対し、同図のbのように、38個の32×36サンプルを出力する。

【0010】サブバンドデータサブサンプリング部2で

は、図5のcに示されているように、単位時間(例えば、1秒)分のサブバンドデータのうち各フレームiの先頭にあるサブバンドデータS<sub>i</sub>(n)(i=0, 1, ..., j-1)を抽出し、図1の各フレームのエネルギー演算部3および第1のメモリ4に入力する。

【0011】以上の動作は、図2では、ステップS1～S9で行われる。ステップS1では、フレーム番号を表すiが0と置かれ、ステップS2ではサブバンド番号を表すnが0と置かれる。ステップS3では、可変長復号部1にて符号化データが可変長復号され、ステップS4ではiフレーム目の先頭のサブバンドデータS<sub>i</sub>(n)が抽出される。次に、ステップS5にて、n=32が成立するか否かの判断がなされ、この判断が否定の時にはステップS6に進んでnに1が加算される。そして、ステップS3に戻って前記と同様の処理が行われる。以上のステップS3～S6の処理が繰り返し行われて、ステップS5の判断が肯定になると、iフレーム目の先頭のサブバンドデータS<sub>i</sub>(n)が抽出されたことになる。

【0012】ステップS5の判断が肯定になると、ステップS7に進み、iに1が加算される。次にステップS8に進み、i=jが成立するか否かの判断がなされる。この判断が否定の時にはステップS2に戻り、再びn=0とされて、再度前記した処理が続行される。以上の処理が繰り返し行われ、ステップS8の判断が肯定になると、i=0～(j-1)フレームの先頭のサブバンドデータS<sub>i</sub>(n)が抽出されたことになり、ステップS9にて、これらのサブバンドデータS<sub>i</sub>(n)は図1の各フレームのエネルギー演算部3および第1のメモリ4に転送されることになる。

【0013】各フレームのエネルギー演算部3では、下記の(1)式に従って各フレームのエネルギーE<sub>i</sub>を計算し、平均エネルギー演算部5および第2のメモリ6に入力する。

【0014】

【数1】

$$E_i = \sum_{n=1}^{32} (S_i(n))^2 \quad \dots\dots (1)$$

S<sub>i</sub>はサブバンドデータ、nはサブバンド番号である。

各フレームのエネルギーE<sub>i</sub>が計算されると、該エネルギーE<sub>i</sub>はステップS10にて平均エネルギー演算部5および第2のメモリ6に転送される。平均エネルギー演算部5では、下記の(2)式に従って入力された各フレームのエネルギーから単位時間間の平均エネルギーA<sub>E</sub>を計算し第1の判定部7に入力する(ステップS11)。

【0015】

【数2】

$$AE = \sum_{i=1}^j E_i \quad \dots\dots (2)$$

jは1秒間のフレーム数である。

第1の判定部7では、入力された単位時間間の音声情報が無音であるのか有音であるのかを、下記の(3)式に従って判定し条件に合う場合には有音であると判定する(ステップS12)。有音である場合には無音である場合に比べて単位時間間の平均エネルギーAEは大きいから、下記の(3)式が成立することになる。

$$【0016】 AE > \alpha \quad \dots(3)$$

ここに、 $\alpha$ は予め定められた第1の閾値である。

【0017】該第1の判定部7において、入力された単位時間間の音声情報が有音であると判断された場合には、第2のメモリ6より各フレームのエネルギー単位時間分を読み出してエネルギー変化率演算部8に入力し(図3のステップS13)、下記の(4)式に従ってエネルギー変化率Cを計算し、第2の判定部9に入力する。一方、無音であると判定された時には、以降の音声、音楽判定処理を終了し、ステップS1に戻る。下式のCは、MPEG符号化データのサブバンドデータから求めた隣り合うフレームの2つのエネルギーの比の単位時間における総和を表している。

$$【0018】$$

【数3】

$$C = \sum_{i=1}^j \left| 10 \log_{10} (E_{i+1}/E_i) \right| \quad \dots\dots (4)$$

音声の時間波形を見ると、単語や音節ごとに波形も変化し、その間は数10m秒にわたって無音となるため、そのスペクトル変化率は、連続波形となる音楽に比べて非常に大きくなる。そこで、第2の判定部9では入力された単位時間の音声情報が音声区間であるか否かを下記の(5)式に従って判定し、条件に合う場合には音声区間と判定し、その区間のタイムコードdを音声区間保持部10に出力する(ステップS14の判断が肯定、ステップS15)。

$$【0019】 C > \beta \quad \dots(5)$$

ここに、 $\beta$ は第2の閾値である。

【0020】一方、音声区間でないと判断された場合には(ステップS14の判断が否定)、第1のメモリ4より各フレームの先頭のサブバンドデータを読み出して平均エネルギー比演算部11に入力する(ステップS16)。

【0021】平均バンドエネルギー比演算部11では、下記の(6)式に従って平均バンドエネルギー比Bmiを計算して第3の判定部12に入力する。

$$【0022】$$

【数4】

$$B_{mi} = \frac{\sum_{n=0}^{k-1} (S_i(n))^2}{\sum_{n=k}^{31} (S_i(n))^2} \quad \dots\dots (6)$$

音声の周波数は、図7(a)に示されているように、一般的に低周波帯域に集中し、一方音楽の周波数は、同図(b)に示されているように、全帯域に分散する傾向がある。換言すれば、音声のサブバンドデータが低周波帯域に集中するのに対して、音楽のサブバンドデータは全帯域にわたって分散する傾向がある。そこで、第3の判定部12では、入力された単位時間間の音声情報が音楽区間であるか否かを下記の(7)式に従って判定し(ステップS17)、条件に合う場合には音楽区間と判定し、その区間のタイムコードeを音楽区間保持部13に出力する(ステップS18)。

$$B_{mi} < \gamma \quad \dots(7)$$

ここに、 $\gamma$ は第3の閾値である。

【0023】以上のように、本実施形態によれば、圧縮符号化されたオーディオの符号化データから、音声区間と音楽区間を区別し、それぞれの区間のタイムコードを音声区間保持部10および音楽区間保持部13のそれぞれに記憶させることができるようになる。

【0024】本発明は、さらに圧縮符号化されていないオーディオ情報の分類に関しても適応できる。その場合の実施形態を以下に示す。

【0025】圧縮符号化されていないオーディオ情報を扱う場合は、図1の可変長復号部1およびサブバンドデータサブサンプリング部2は高速フーリエ変換部(以下、FFT変換部と呼ぶ)に置き換えられる。元のオーディオ情報からこのFFT変換部において、図6にあるようにFFT変換を行い、単位時間分の周波数データを抽出する。今、該単位時間を1秒とすると、元のオーディオ信号pからサンプリングした2048個のサンプルをFFT変換し、それを時間的にサンプルをずらしながら38回繰り返して、合計2048×38個のFFTデータを単位時間分の周波数データとしている。

【0026】その後、各フレームのエネルギー演算部、平均エネルギー演算部、エネルギー変化率演算部、および平均バンドエネルギー比演算部で、それぞれ下記の(8)式、前記(2)式、(4)式、および下記の(9)式に従ってそれぞれ各フレームのエネルギーEi、平均エネルギーAE、エネルギー変化率C、平均バンドエネルギー比Bmiを計算し、第1の判定部7、第2の判定部9、第3の判定部12にてそれぞれ無音/有音の判定、音声の判定、音楽の判定を行う。

$$【0027】$$

【数5】

$$E_i = \sum_{n=1}^{2048} (F_i(n))^2 \quad \dots\dots (8)$$

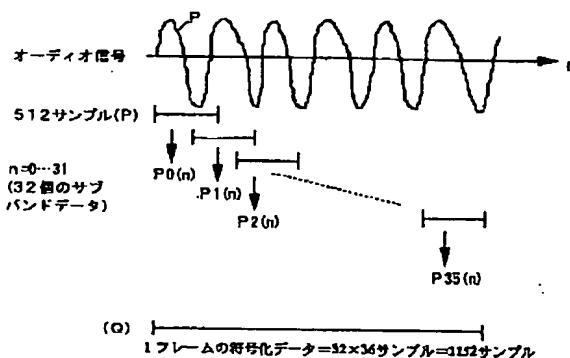
$$B_{mi} = \frac{\sum_{n=1}^{2048} (F_i(n))^2}{\sum_{n=1}^{2048} (F_i(n))^2} \quad \dots\dots (9)$$

## 【0028】

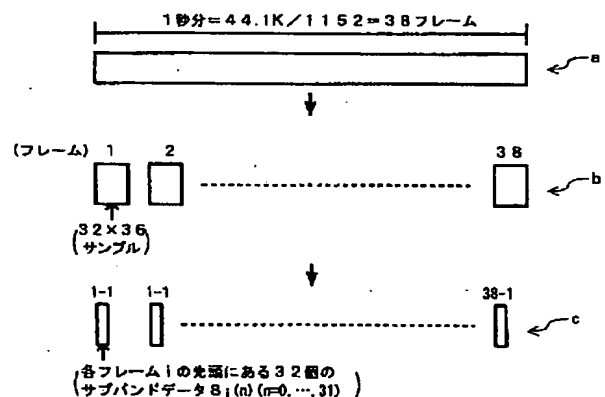
【発明の効果】以上説明したように、本発明によれば、圧縮符号化された音声データから符号化データ上でオーディオ情報を音声区間と音楽区間に分類することが可能になるという効果がある。

【0029】なお、本発明を実際に動作させたところ、次のような結果が得られた。すなわち、MPEG1レイヤ2で符号化された合計90分のニュース番組と音楽番組のオーディオビットストリームを用いて、1秒毎の音声区間と音楽区間の分類を行った。音声区間は背景に音楽などがなく音声のみが含まれる区間を対象とし、音楽区間は音声の有無にかかわらず楽器演奏がある区間を対象とした。音声区間の検出に関しては、89.4%、音楽区間に関しては79.3%の検出率を得ることができ、音声区間の検出に関しては実用レベルの検出率を得ることができた。また、本発明によれば、圧縮符号化されていないオーディオ情報の分類に関しても、簡単に、

【図4】



【図5】



音声区間と音楽区間に分類することが可能になるという効果がある。

## 【図面の簡単な説明】

【図1】 本発明の一実施形態のオーディオ情報分類装置の構成を示すブロック図である。

【図2】 図1の制御装置の動作を説明するためのフローチャートである。

【図3】 図2の続きの動作を説明するためのフローチャートである。

10 【図4】 MPEGオーディオ符号化データの構造を説明するための図である。

【図5】 図1のサブバンドデータサブサンプリング部の動作を説明するための図である。

【図6】 符号化されていないオーディオ情報の周波数データの抽出方法を説明するための図である。

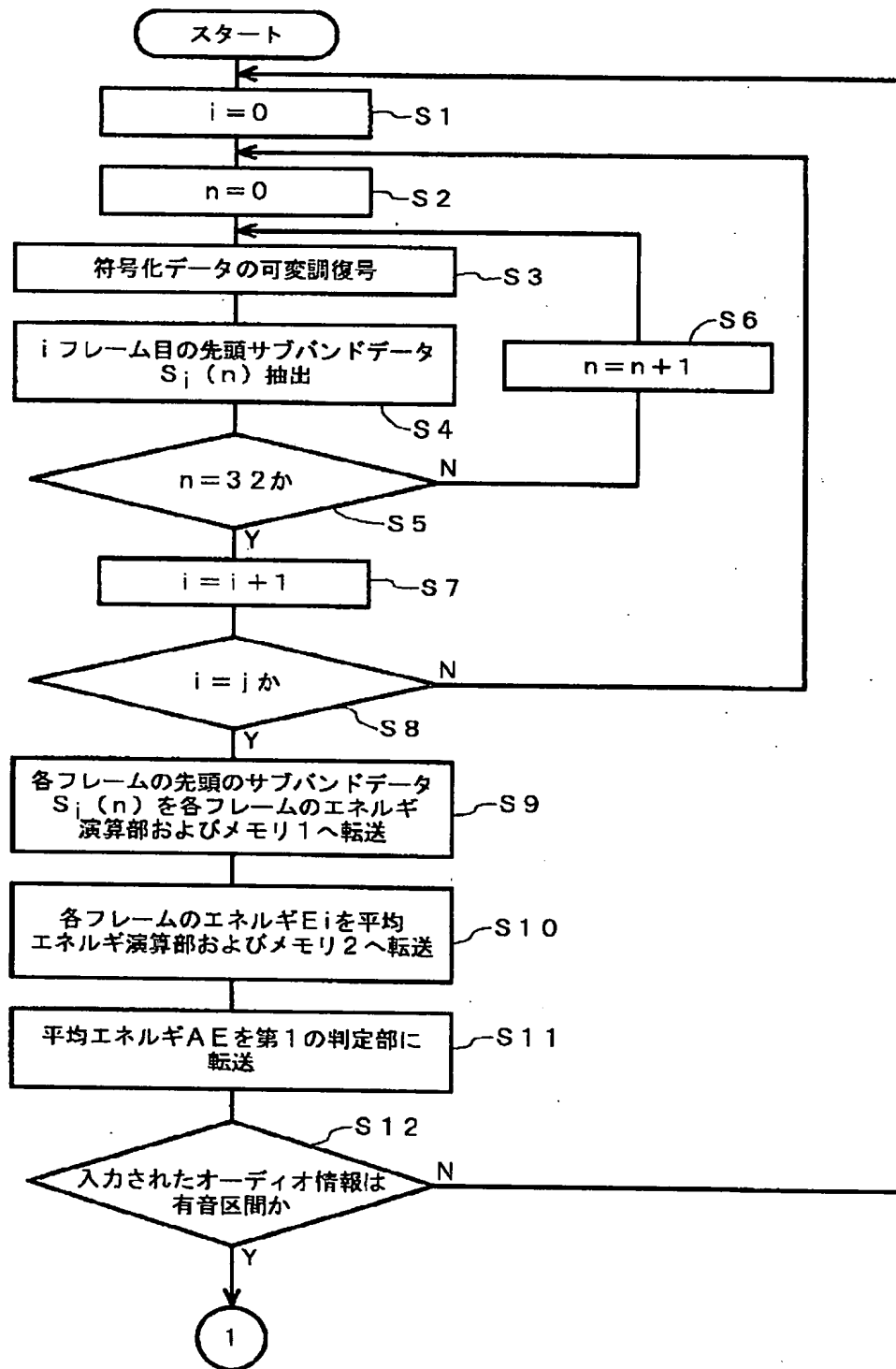
【図7】 音声と音楽の周波数分布の傾向を示す図である。

## 【符号の説明】

1…可変長復号部、2…サブバンドデータサブサンプリング部、3…各フレームのエネルギー演算部、4…第1のメモリ、5…平均エネルギー演算部、6…第2のメモリ、7…第1の判定部、8…エネルギー変化率演算部、9…第2の判定部、10…音声区間保持部、11…平均バンドエネルギー比演算部、12…第3の判定部、13…音楽区間保持部。

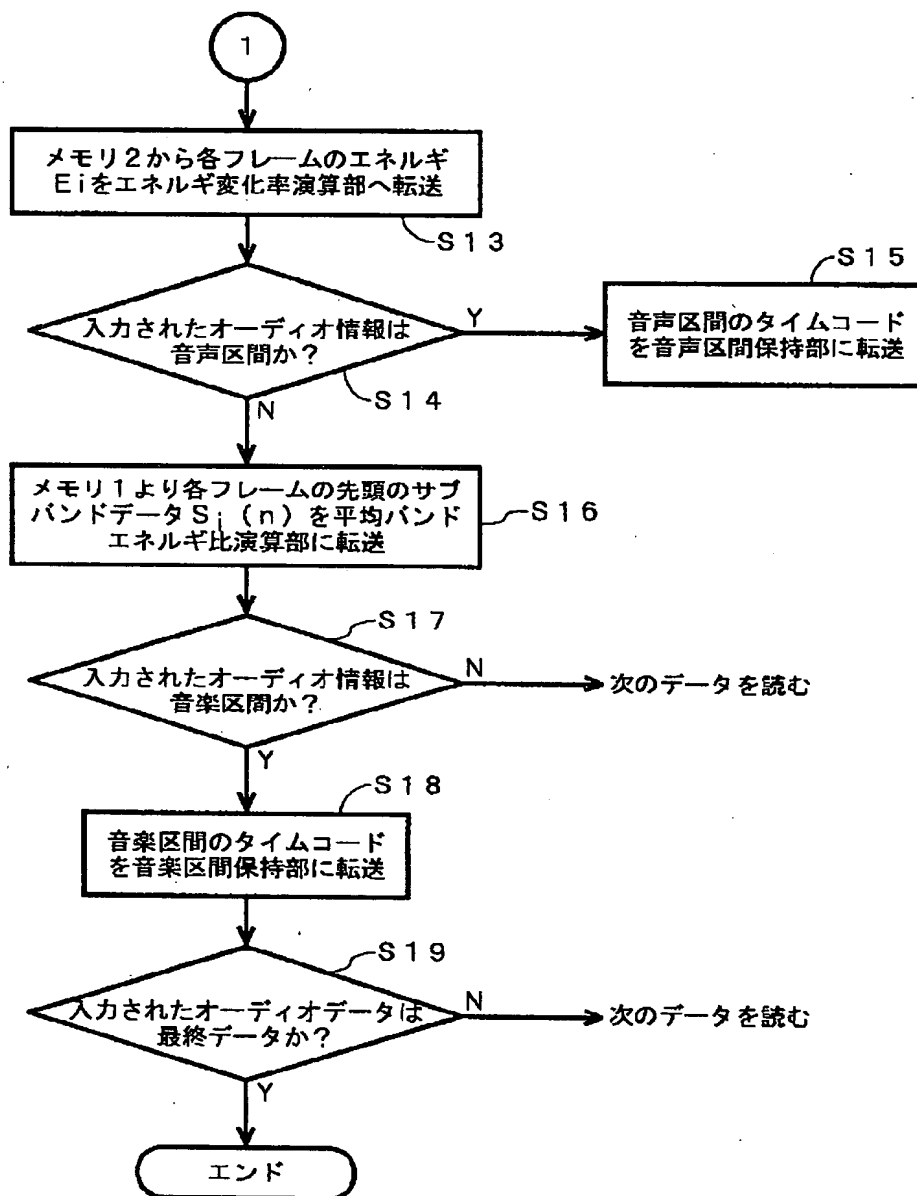
[illegible]

【図2】

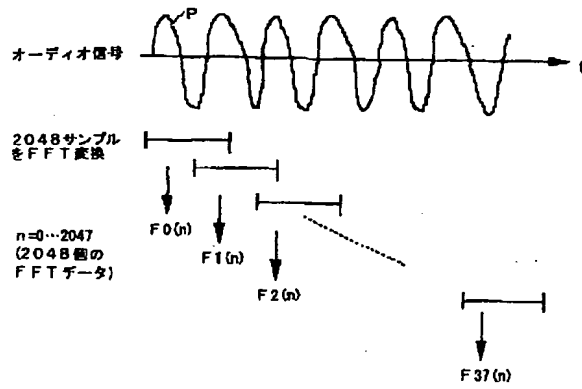




【図3】



【図6】



【図7】

